How to improve robustness in Kohonen maps and display additional information in Factorial Analysis: application to text mining

Nicolas Bourgeois, Marie Cottrell, Benjamin Deruelle, Stéphane Lamassé, Patrick Letrémy

How to improve robustness in Kohonen maps and visualization in Factorial Analysis: application to text mining

Nicolas Bourgeois[a], Marie Cottrell[a], Benjamin Déruelle[b], Stéphane Lamassé[b], Patrick Letrémy[a]

[a]*SAMM - Université Paris 1 Panthéon-Sorbonne 90, rue de Tolbiac, 75013 Paris, France*
nbourgeo@phare.normalesup.org,marie.cottrell,patrick.letremy@univ-paris1.fr
[b]*PIREH-LAMOP - Université Paris 1 Panthéon-Sorbonne 1, rue Victor Cousin, Paris, France*
benjamin.deruelle,stephane.lamasse@univ-paris1.fr

Abstract

This article is an extended version of a paper presented in the WSOM'2012 conference [1]. We display a combination of factorial projections, SOM algorithm and graph techniques applied to a text mining problem. The corpus contains 8 medieval manuscripts which were used to teach arithmetic techniques to merchants.

Among the techniques for Data Analysis, those used for Lexicometry (such as Factorial Analysis) highlight the discrepancies between manuscripts. The reason for this is that they focus on the deviation from the independence between words and manuscripts. Still, we also want to discover and characterize the common vocabulary among the whole corpus.

Using the properties of stochastic Kohonen maps, which define neighborhood between inputs in a non-deterministic way, we highlight the words which seem to play a special role in the vocabulary. We call them fickle and use them to improve both Kohonen map robustness and significance of FCA visualization. Finally we use graph algorithmic to exploit this fickleness for classification of words.

Introduction

Historical Context

One approach to understand the evolution of science is the study of the evolution of the language used in a given field. That is why we would like to pay attention to the vernacular texts dealing with practical arithmetic and written for the instruction of merchants. Such texts are known since the XIII$^{\text{th}}$ century, and from that century onwards, the vernacular language appears more and more as the medium of practical mathematics.

Treaties on arithmetical education were therefore mostly thought and written in local languages, (they were written not only in French but also in Italian, Spanish, English and German). In this process, the XV$^{\text{th}}$ century appears as a time of exceptional importance because we can study the inheritance of two hundred years of practice. For the authors of these texts, the purpose was not only to teach merchants, but also to develop knowledge in vernacular language. Their books were circulated far beyond the shopkeepers' world, to the humanists' circles for example.

An objective of historical research: the study of specialized languages

The work previously done by historians [2] consisted in the elaboration of a dictionary of the lexical forms found in all the treaties, in order to identify the different features of the mathematical vernacular language at that time. This being done, we have worked on the contexts of some especially important words in order to understand the lexicon in all its complexity. In other words, we would like to determine the common language that forms the specialized language beyond the specificity of each text.

Manuscripts and **Title**	**Date**	**Author**	**Number of occurrences**	**Number of words**	**Hapax**
Bib. nat. Fr. 1339	ca.1460	anonyme	32077	2335	1229
Bib. nat. Fr. 2050	ca.1460	anonyme	39204	1391	544
Cesena Bib. Mal. S-XXVI-6, *Traicté de la praticque*	1471?	Mathieu Préhoude?	70023	1540	635
Bibl. nat. Fr. 1346, Commercial appendix of *Triparty en la science des nombres*	1484	Nicolas Chuquet	60814	2256	948
Méd. Nantes 456	ca.1480-90	anonyme	50649	2252	998
Bib. nat. Arsenal 2904, *Kadran aux marchans*	1485	Jean Certain	33238	1680	714
Bib. St. Genv. 3143	1471	Jean Adam	16986	1686	895
Bib. nat. Fr. Nv. Acq. 10259	ca.1500	anonyme	25407	1597	730

Table 1: Corpus of texts and main lexicometric features. The number of occurences is the total number of words including repetitions, the number of words is the number of distinct words, Hapax are words appearing once in a text.

Outline of this work

Among the techniques for Data Analysis, those used for Lexicometry (such as Factorial Analysis) highlight the discrepancies between manuscripts. The reason for this is that they focus on the deviation from the independence between words and manuscripts. Still, we also want to discover and characterize the common vocabulary among the whole corpus. That is why we introduce a new tool, which combine the properties of Factorial Correspondence Analysis and Stochastic Self-Organizing Maps. That leads to the definition of fickle pairs and fickle words. Fickle words can be seen as this common vocabulary we are looking for, and prove themselves to be a good basis for a new visualization with the help of graph theory.

In part 1, we first focus on the definition of the corpus: the texts, the pre-processing, and the protocol which is traditionally used in Humanities and Social Sciences to handle such data. Then, (part 2), we design the tools : 'fickle pairs' and 'fickle words', robust Kohonen Maps, improved FCA, graphs of relations between words based on fickleness. We explain the algorithms involved and display the results on the corpus. Finally (part 3), we give a brief analysis and comments on these results.

1. Text, Corpus and protocol

In order to delimit a coherent corpus among the whole European production of practical calculation education books, we have chosen to pay attention to those treaties which are sometimes qualified as commercial (*marchand* in French) which have been written in French between 1415 and about 1500. Note that this corpus has already been studied by [3], [4] and [5]. In this way, our corpus follows the rules of the discourse analysis: homogeneity, contrastiveness and diachronism. For further explanation about texts, methodology and purpose of the analysis see [2], for further explanation about the corpus [6], for wider explication about analysis see [7], [8].

It contains eight treaties on the same topic, written in the same language and by different XV$^{\text{th}}$ century authors. The following Table 1 describes some elements of the lexicometric characteristics of the corpus and shows how non balanced it is.

1.1. Humanities and Social Sciences traditional protocol

Traditionally on this kind of textual data, researchers in Humanities and Social Sciences work on statistical specificity and contextual concordances, since they allow an easy discovery of the

major lexical splits within the texts of the corpus, while remaining close to the meanings of the different forms.

Then, the factorial and clustering methods, combined with co-occurrences analysis (see [9]) help us to cluster the texts without breaking the links with semantic analysis.

However, such a method of data processing requires a preliminary treatment of the corpus, the lemmatization [10]. It consists in gathering the different inflected forms of a given word as a single item. It allows us to work at many different levels of meaning, depending upon the granularity adopted: forms, lemma, syntax.

We can justify this methodological choice here by its effect on the dispersion of the various forms which can be linked to the same lemma, a high degree of dispersion making the comparison between texts more difficult. It must also be remembered that in the case of medieval texts, this dispersion is increased by the lack of orthographic norms. In our case, this process has an important quantitative consequence on the number of forms in the corpus, which declines from 13516 forms to 9463, a reduction of some 30%.

This process has been achieved with a particular attention to the meaning of each word in order to suppress ambiguities: a good example is the French word *pouvoir* which can be a verb translated by "can" or "may", and which is also a substantive translated by "power".

Finally, to realize a clustering of the manuscripts, we have only kept the 219 words with highest frequencies. The set of words selected that way for text classification relates to mathematical aspects, such as operations, numbers and their manipulations, as well as to didactic aspects. Their higher frequencies reflect the fact that they are the language of the mathematics as they appear to be practiced in these particular texts.

Thus, in what follows, the data are displayed in a contingency table T with $I = 219$ rows (the words) and $J = 8$ columns (the manuscripts) so that the entry $t_{i,j}$ is the number of occurrences of word i in manuscript j.

1.2. Use of Factorial Correspondence Analysis (FCA)

Factorial Correspondence Analysis is one of the factorial methods which consist in applying an orthogonal transformation to the data, to supply the user with simplified representation of high-dimensional data, as defined in [11]. The most popular of these factorial methods is the Principal Component Analysis, which deals with real-valued variables and supplies for example the best two-dimensional representation of high-dimensional dataset, by retaining the first two eigenvectors of the covariance matrix.

Factorial Correspondence Analysis (see [12] or [13]) is a variant of Principal Component Analysis, designed to deal with categorical variables. Let us consider two categorical variables with respectively I and J items and the associated contingency table T where entry $t_{i,j}$ is the number of co-occurrences of item i for the row variable and item j for column variable. The rows and the columns are scaled to sum to 1 and normalized in order to be treated simultaneously, by defining

$$t_{i,j}^{norm} = \frac{t_{i,j}}{\sqrt{\sum_i t_{i,j} \sum_j t_{i,j}}}. \tag{1}$$

To achieve the FCA, two Principal Component Analysis are made over the normalized table T^{norm} (providing a representation of the rows) and its transposed table (providing a representation of the columns). The main property of FCA is that both representations can be superposed, since their principal axes are strongly correlated. The proximity between items is significant, regardless they stand for row items or column items, except in the center of the map.

In our case, the rows are the words ($I = 219$) and the columns are the manuscripts ($J = 8$). Figures 2 and 3 show the projection of the data on the first four factorial axes.

The first two factors (43.94% of the total variance) show the diversity of the cultural heritages which have built the language of these treaties. The first factor (25.03%) discriminates between the university legacy on the right, and the tradition of mathematical problems on the left.

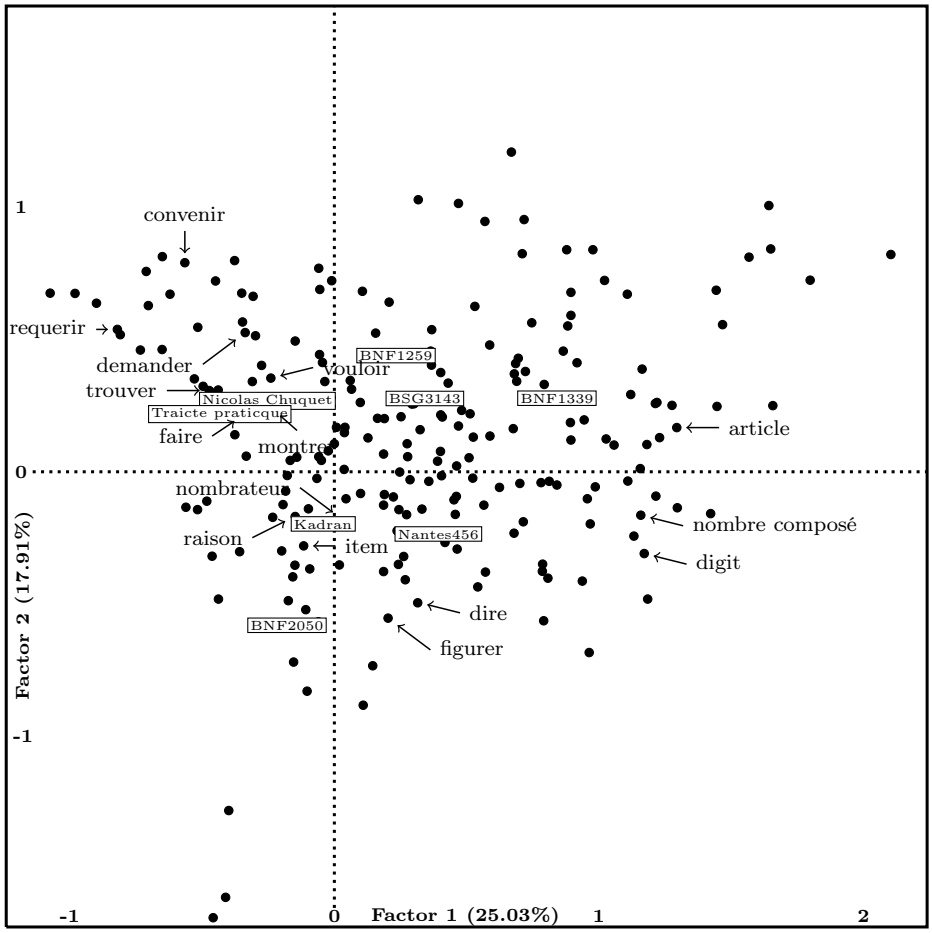

Figure 2: Projection on the first two factors of the FCA. The eight texts appear in frames, a few words are displayed while the remaining are simply figured by dots, for the sake of readability.

On the left, we can observe a group whose strong homogeneity comes from its orientation towards mathematical problems (*trouver* that is to say "to find", *demander* which we can translate as "to ask") and their iteration (*item, idem*). That vocabulary can be found most often in both the appendix of *Triparty en la science des nombres* (Nicolas Chuquet) and *Le Traicte de la praticque*. Furthermore, there are more verbal forms on this side of the axis than on the other. And we can find verbs like *requerir* which means "to require", *convenir* "to agree", *faire* "to do". Some of them are prescriptive, as *devoir* "to have to" or *vouloir* "to want" for example, while others introduce examples, as *montrer* "to show". All these texts contain a lot of mathematical problems and in a way are practical. On the right, the texts of BnF. fr. 1339 and Med. Nantes 456 are clearly more representative of the university culture, containing latin words sequences.

The second axis (17.91% of the variance) is mostly characterized by the manuscript of BNF. fr. 2050 and also by *Kadran aux marchans*. It displays words of Italo-Provencal origin, like *nombrateur* which refers to the division's numerator. Designations of the fraction and operation of division take a significant part of the information while the most contributory words (for ex. *figurer* "to draw") allow us to examine another dimension of these works: the graphical representation as a continuation of writing.

The following factors 3, 4 and further show the Lexicon that seems to be more related to the singularity of some manuscripts. The importance of Nicolas Chuquet inertia of factors 3 and 4 singles out this book on the plane (see Figure 3) in relation to the rest of the corpus.

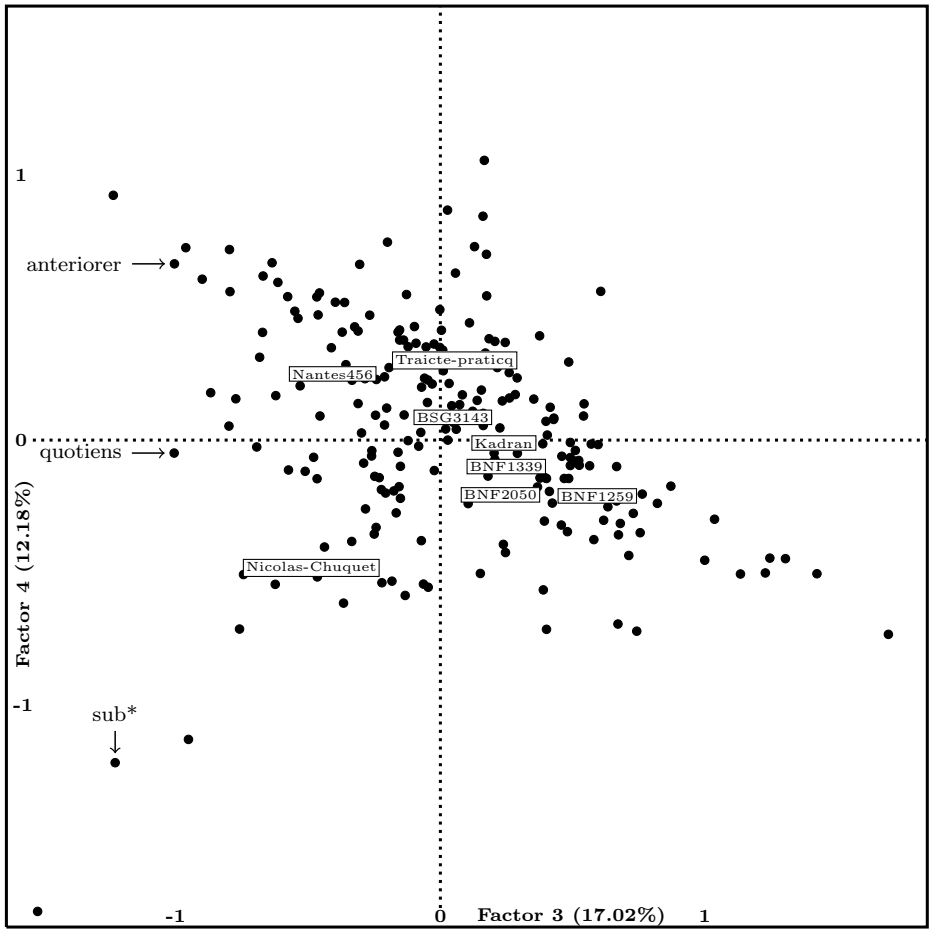

Figure 3: Projection on third and fourth factors of the FCA.

With the manuscript of Nantes 456, at left, factor 3 highlights a vocabulary of some technical accuracy, in any case rare in the rest of the corpus, like *quotiens* "quotient", *anteriorer* "to put before". At right, there is a very diversified vocabulary, associated to manuscrit 10259, which is a well organized compilation of a copy of *Kadrans aus marchans* and of a lot of problems whose origin has not been fully identified.

Correspondence Analysis displays the particularities of each text, but leaves untouched some more complex elements of the data. For instance, we have to see the third factor to understand that the *Triparty en la science des nombres* (Nicolas Chuquet) and the *Traicte de la praticque* use different university mathematical cultures. These two treaties are not only copying university algorithms as they were taught at university at that time, they have their own originality.

Moreover, we cannot assert that the words which appear in the center of the graph represent a 'common vocabulary': as a matter of fact, we should analyze all the successive factors in order to build the list of words constituting the 'common vocabulary'. It is a very cumbersome task.

1.3. Kohonen Maps

SOM-based algorithms were very often used for text mining purpose. Oja and Kaski's seminal book [14] provides a lot of examples on this field. A major tool for that purpose is WEBSOM method and software[1], as defined for instance in [15, 16, 17]. Other important papers (among

[1] See `http://websom.hut.fi/websom/`

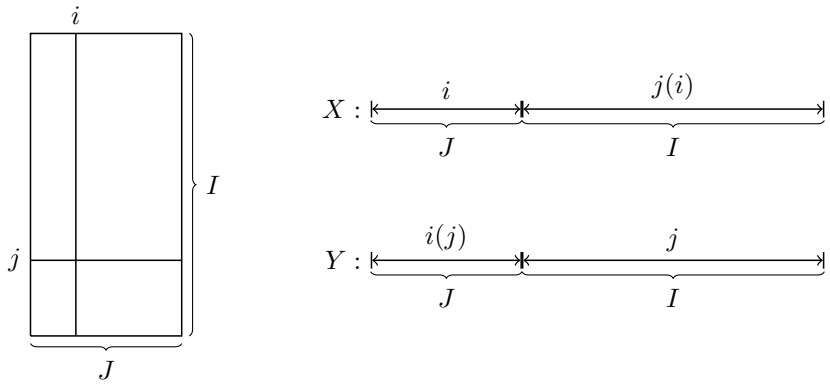

Figure 4: Building of the extended, symmetrized table in the KORRESP algorithm

hundreds) are [18, 19, 20].

Most of them look for classification and clustering using keywords, put in evidence the main features, associate documents with their most characteristic words, to define proximity in order to define clusters and hierarchies between documents. Techniques such as WEBSOM are especially designed to deal with massive documents collections.

Our purpose is very different, since we have very few documents and since we look for a subset of words which are not "specific" of some manuscript, but contrarily belong to a common vocabulary.

Factorial Correspondence Analysis (FCA) suffers from some limitations as explained in section1.2. To overcome this, we use a variant of the SOM algorithm which deals with the same kind of data, i. e. a contingency table. This variant of SOM was defined in [21] or [22] and we refer to it as KORRESP algorithm. Let us recall this definition.

Tha data are displayed as explained in fourth paragraph of section 1.2 in a contingency $I = 219$ by $J = 8$ table. The data are normalized applying equation (1), exactly in the same way as for Factorial Correspondence Analysis. The normalized contingency table is denoted by T^{norm} where:

$$t_{i,j}^{norm} = \frac{t_{i,j}}{\sqrt{\sum_i t_{i,j} \sum_j t_{i,j}}}.$$

We consider a Kohonen map, and associate to each unit u a code-vector C_u with $(J + I)$ components. The first J components evolve in the space of the rows (the words), while the last I components belong to the space of the columns (the manuscripts).

Let us denote

$$C_u = (C_J, C_I)_u = (C_{J,u}, C_{I,u}), \tag{2}$$

to put in evidence the structure of the code-vector C_u.

We use the SOM algorithm as a double learning process, by alternatively drawing a T^{norm} row (a word) and a T^{norm} column (a manuscript).

When we draw a row i, we associate the column $j(i)$ that maximizes the coefficient $t_{i,j}^{norm}$, so:

$$j(i) = \arg\max_j t_{i,j}^{norm} = \arg\max_j \frac{t_{i,j}}{\sqrt{\sum_i t_{i,j} \sum_j t_{i,j}}} \tag{3}$$

that maximizes the conditional probability of j given i. We then create an extended $(J + I)$ - dimensional row vector $X = (i, j(i)) = (X_J, X_I)$. See Figure 1.3.

Subsequently, we look for the closest of all the code vectors, in terms of the Euclidean distance restricted to the first J components. Note u_0 the winning unit. Next we move the code-vector of the unit u_0 and its neighbors towards the extended vector $X = (i, j(i))$, as per the customary Kohonen law. Let us write down the formal definition:

$$u_0 = \arg\min_u \|X_J - C_{J,u}\| \quad (4)$$

$$C_u^{new} = C_u^{old} + \epsilon\sigma(u, u_0)(X - C_u^{old}) \quad (5)$$

where ϵ is the adaptation parameter (positive, decreasing with time), and σ is the neighborhood function, such that $\sigma(u, u_0) = 1$ if u and u_0 are neighbour in the Kohonen network, and $\sigma(u, u_0) = 0$ if not.

The reason to associate a row and a column in such a way is to keep the row-column associations which are realized in classical FCA by the fact that the principal axes of both Principal Component Analysis are strongly correlated.

The procedure is the same when we draw a column j with dimension I (a column of T^{norm}). We associate the row $i(j)$ that maximizes the coefficient $t_{i,j}^{norm}$, so:

$$i(j) = \arg\max_i t_{i,j}^{norm} = \arg\max_i \frac{t_{i,j}}{\sqrt{\sum_i t_{i,j} \sum_j t_{i,j}}} \quad (6)$$

that maximizes the conditional probability of i given j. We then create an extended $(J + I)$-dimensional column vector $Y = (i(j), j) = (Y_J, Y_I)$.

We then seek the code-vector that is the closest, in terms of the Euclidean distance restricted to the last I components. Let v_0 be the winning unit. Next we move the code-vector of the unit v_0 and its neighbors towards the extended vector $Y = (i(j), j)$, as per the customary Kohonen law. Let us write down the formal definition:

$$v_0 = \arg\min_v \|Y_I - C_{I,u}\| \quad (7)$$

$$C_v^{new} = C_v^{old} + \epsilon\sigma(v, v_0)(Y - C_v^{old}) \quad (8)$$

where ϵ and σ are defined as before.

This two-steps computation carries out a Kohonen classification of the rows (the words), together with a classification of the columns, maintaining all the while the associations of both rows and columns.

We can sum up the definition of the KORRESP algorithm

- *normalization* of the rows and of the columns in the way as in FCA computation,
- *definition* of an extended data table by associating to each row the most probable column and to each column the most probable row,
- *simultaneous classification* of the rows and of the columns onto a Kohonen map, by using the rows of the extended data table as input for the SOM algorithm.

After convergence of the training step, the items of the rows and of the columns are simultaneously classified. In our example, one can see proximity between words, between texts, between words and texts. It is the same goal as in Factorial Correspondence Analysis. The advantage is that it is not necessary to examine several projection planes: the whole information can be read on the Kohonen Map.

We display below (Figure 5) the SOM map which simultaneously represents the words and the texts. For this map as for all the remaining of this paper we use the online algorithm, a 10×10 grid and the following simple neighborhood function: 1 for the eight (fewer if we are along one edge of the map) nodes adjacent to the selected one and 0 for the others.

One can observe that the interpretation (see Figure 6) is very similar to the interpretation that could be done from the Factorial Correspondence Analysis projections. But, as an example of the robustness problem, we can compare two different Kohonen maps (in Figures 5 and 7) and the respective positions of the words *raison* "reason" and *dire* "to say", very far from each other in the first map while neighboring in the second one.

minutes super*		notes	calculer cubic	fois mettre **BNF10259**		contraire depenser falloir meme racine	aller donc ensuivre savoir	multiplier	regle venir
gecter					dessous	barrater demi	somme	voir	assembler
	BSG3143	parteur	defaillir duplation mediation nommer numeration senestre	semblables		circulaires demeurer derenier disaine ecrire entendre formes nombrateur oter prouver	entrer laisser rien	emprunter figure regarder	figure de non rien fraction muer rayes retenir
notables		denomi-nations multipli-cateur nominateur ordonne	abaisser comptes endroit proposer	anteriorer diminution enseignement enseigner moyen signifiant surplus trancher	possible reduire	difficile progression repondre	avaluer	faillir	monter partiteur
bref gectons multipli-cation pratique seulement	generale latin nombrer proportion reduction unite	soustraction	denomi-nateur	entier	ajoutement		remotion **Kadran**		**BNF2050**
chose	nulle	ensemble partie	**Nantes456**				partement valoir		etre
arithmetique compter preuve tenir	cubbement destre digit diviser diviseur division lignes nombre composequer	**DIRE** figurer poser science			abreger lever precedent quotiens	numerateur	moindre nombre **Nicolas Chuquet**	partir plus **RAISON**	
egalir egaliser especes question total traiter	mesurer	grand	apparaitre position soustraire	ajouter	leurs prendre quant quantefois reponse trouver		part	commun Item	devoir droit exemple reste rester
algorisme article carrees cercle envient ligne pair sain **BNF1339**	addition chiffre former	double doubler moitie	appeler donner maniere pouvoir se	bailler demander mises nomper pareillement vouloir		egale faire montrer necessaire romp selon			
cautelle	demontrer dessus	garder regle de trois	aliquot composer corps moins sub* toutefois	partant plaisir proportionel-lement requerir residu	appartenir convenir demande difference egaulx maieur millions rate survendre tant		naturel roupt **Traicte praticque**		fausse

Figure 5: Example of Kohonen Map. Manuscripts are in bold. Notice that *raison* (9,7) and *dire* (3,7) are far apart from each other.

The Kohonen algorithm is stochastic, and it can happen that several runs get different results, and that these differences can be troublesome. Hence the idea to introduce repetitions of the runs to separate stable and robust results from purely stochastic behavior. In the following, we study the variability of the maps which provides new information.

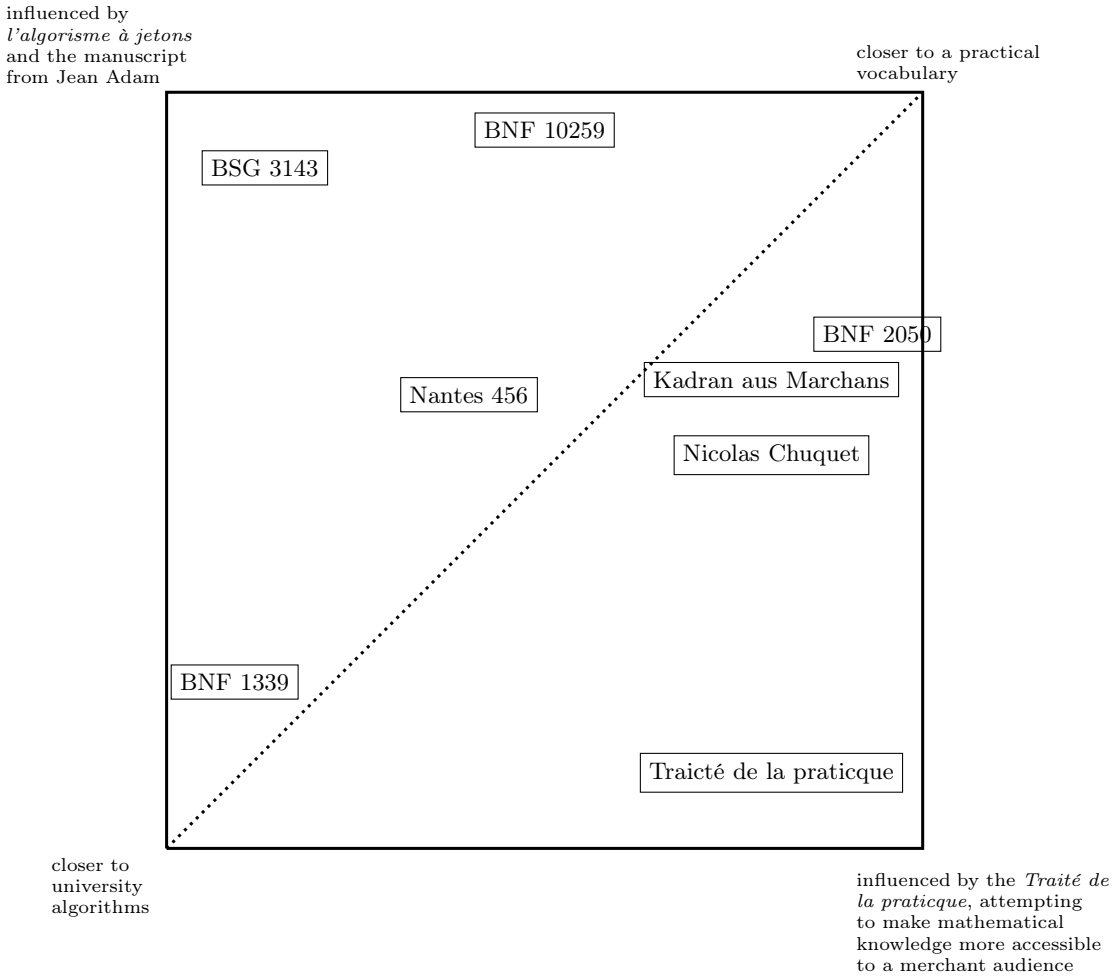

Figure 6: Interpretation of the Kohonen map (Figure 5), the diagonal opposes university and practical vocabularies

2. Getting extra information through the extraction of fickle words

In its classical presentation [23, 21], the SOM algorithm is an iterative algorithm, which takes as input a dataset $\mathbf{x}_i, i \in \{1, \ldots, N\}$ and computes code-vectors $\mathbf{m}_u, u \in \{1, \ldots, U\}$ which define the map.

We know that self-organization is reached at the end of the algorithm, which implies that close data in the input space have to belong to the same class or to neighboring classes, that is to say that they are projected on the same prototypes or on neighboring prototypes on the map. In what follows, we call neighbors data that belong either to the same unit or to two adjacent units. But the reciprocal is not exact: for a given run of the algorithm, two given data can be neighbors on the map, while they are not in the input space. That drawback comes from the fact that there is no perfect fit between a two-dimensional map and the data space (except when the intrinsic dimension is exactly 2). As we just notice, since the SOM algorithm is a stochastic one, the resulting maps can be different from one run to another. How to overcome this difficulty?

In fact, we can use this drawback to improve the interpretation and the analysis of relations between the studied words. Our hypothesis is that the repetitive use of this method can help us to identify words that are strongly attracted/repulsed and also fickle pairs.

cautelle latin nombrer pair	demontrer	dessus **BNF1339**	addition traiter	compter double doubler somme		devoir droit multiplier rien	barrater contraire depenser falloir meme racine		calculer cubic fois mettre **BNF10259**
unite	article chiffre former			demeurer preuve	remotion			dessous	
duplation mediation		carrees cercle envient ligne sain		progression tenir		faillir		formes oter	demi
anteriorer cubbement destre digit diviser diviseur division lignes moyen nombrecompose signifiant	egalir especes querir	algorisme egaliser total	arithmetique circulaires mesurer question	garder laisser				disaine ecrire entendre nombrateur possible	prouver reduire
ajoutement apparaitre diminution ensemble partie precedent proposer semblables	science soustraction	**DIRE** figurer poser	ajouter moitie position pouvoir **RAISON** se soustraire	moins regle de trois	partement	fausse	exemple	avaluer derenier difficile entrer repondre Kadran	
Nantes456			assembler grand valoir	appeler composer donner	aller bailler mises	convenir demande demander donc nomper pareillement	partant		
	voir	pratique regle	chose			ensuivre plaisir requerir residu tant	faire	monter	emprunter figure figure de non rien muer partiteur rayes regarder retenir **BNF2050**
denominateur entier	fraction	bref generale multiplication	nulle proportion	part seulement	montrer vouloir	leurs necessaire quant savoir selon trouver	rester venir	reste	
comptes endroit enseignement enseigner nommer senestre surplus trancher	parteur	nominateur notes	gectons ordonne	egale	romp roupt			abreger commun item lever nombre partir plus prendre reponse	
abaisser defaillir denominations multiplicateur numeration reduction		gecter minutes super **BSG3143**	notables	appartenir difference egaulx maieur millions rate survendre	naturel	etre **Traicte praticque**		maniere moindre numerateur quantefois quotiens	aliquot corps proportionel-lement sub toutefois **Nicolas Chuquet**

Figure 7: Another example of Kohonen Map. This time, *raison* (4,5) and *dire* (3,5) are neighbors.

2.1. Neighborhood and robustness of information on Kohonen maps

We address the issue of computing a reliability level for the neighboring (or no-neighboring) relations in a SOM map. More precisely, if we consider several runs of the SOM algorithm, for a given size of the map and for a given data set, we observe that most of pairs are almost always neighbors or always not neighbors. But there are also pairs whose associations look random. These pairs are called *fickle* pairs. This question was addressed by [24] in a bootstrap frame.

According to their paper, we can define: $NEIGH^l_{i,j} = 0$ if x_i and x_j are not neighbors in the l-th run of the algorithm, and $NEIGH^l_{i,j} = 1$ if x_i and x_j are neighbors in the l-th run of the algorithm, where (x_i, x_j) is a given pair of data, l is the number of the observed runs of the SOM algorithm.

Then $Y_{i,j} = \sum_{l=1}^{L} NEIGH^l_{i,j}$ is the number of times when the data x_i and x_j are neighbor for

L different, independent runs. The stability index $\mathcal{M}_{i,j}$ is defined as the average of $NEIGH_{i,j}$ over all the runs ($l = 1, \ldots, L$), i. e.

$$\mathcal{M}_{i,j} = \frac{\sum_{l=1}^{L} NEIGH_{i,j}^{l}}{L} = \frac{Y_{i,j}}{L}. \tag{9}$$

The next step is to compare it to the value it would have if the data x_i and x_j were neighbors by chance in a completely random way.

So we can use a classical statistical test to check the significance of the stability index $\mathcal{M}_{i,j}$. Let U be the number of units on the map. If edge effects are not taken into account, the number of units involved in a neighborhood region (as defined here) is 9 in a two-dimensional map. So for a fixed pair of data x_i and x_j, the probability of being neighbors in a random way is equal to $9/U$ (it is the probability for x_j to be a neighbor of x_i by chance once the class x_i belongs to is determined).

As $Y_{i,j} = \sum_{l=1}^{L} NEIGH_{i,j}^{l}$ is the number of times when the data x_i and x_j are neighbor for L different, independent runs, it is easy to see that $Y_{i,j}$ is distributed as a Binomial distribution with parameters L and $9/U$.

Using the classical approximation of Binomial Distribution by a Gaussian one (L is large and $9/U$ not too small), we can build the critical region of the test of null hypothesis H_0 "x_i and x_j are neighbors by chance" against hypothesis H_1: " the fact that x_i and x_j are neighbors or not is significant".

We conclude that the critical region for a test level of 5% based on $Y_{i,j}$, is

$$]-\infty, L\frac{9}{U} - 1.96\sqrt{L\frac{9}{U}(1-\frac{9}{U})}[\quad \bigcup \quad]L\frac{9}{U} + 1.96\sqrt{L\frac{9}{U}(1-\frac{9}{U})}, +\infty[\tag{10}$$

For the frequency (i.e. the stability index) $\mathcal{M}_{i,j} = Y_{i,j}/L$, the critical region is

$$]-\infty, \frac{9}{U} - 1.96\sqrt{\frac{9}{UL}(1-\frac{9}{U})}[\quad \bigcup \quad]\frac{9}{U} + 1.96\sqrt{\frac{9}{UL}(1-\frac{9}{U})}, +\infty[\tag{11}$$

To simplify the notations, , let us put

$$A = \frac{9}{U} \text{ and } B = 1.96\sqrt{\frac{9}{UL}(1-\frac{9}{U})}. \tag{12}$$

Then, practically, for each pair of words (x_i, x_j), we compute the index $\mathcal{M}_{i,j} = Y_{i,j}/L$, and apply the following rule:

- if their index is greater than $A + B$, they are almost always neighbors in a significant way, the words attract each other.
- if their index is comprised between $A - B$ and $A + B$, their proximity is due to randomness, they are a fickle pair.
- if their index is less than $A - B$, they are almost never neighbor, the words repulse each other.

2.2. Identification of fickle pairs

We run KORRESP L times and store the result in a matrix $\mathcal{M}$ of size $(N + p) \times (N + p)$. The value stored in a given cell i, j is the proportion of maps where i and j are neighbors.

Table 8 displays an example of the first nine rows and columns of such a matrix. We have highlighted with colors three different situations. According to the theoretical study mentioned above:

	abaisser	abreger	addition	ajoutement	ajouter	algorisme	aliquot	aller	anteriorer
abaisser	1	0	0.025	0.275	0	0.05	0	0	0.525
abreger	0	1	0	0	0.25	0	0.325	0	0.025
addition	0.025	0	1	0	0	0.875	0	0.05	0
ajoutement	0.275	0	0	1	0.025	0	0	0.025	0.7
ajouter	0	0.25	0	0.025	1	0.025	0.15	0.125	0
algorisme	0.05	0	0.875	0	0.025	1	0	0	0
aliquot	0	0.325	0	0	0.15	0	1	0.025	0
aller	0	0	0.05	0.025	0.125	0	0.025	1	0
anteriorer	0.525	0.025	0	0.7	0	0	0	0	1

Table 8: Frequency of neighborhood matrix (excerpt)

	abaisser	ajoutement	anteriorer	abreger	ajouter	aliquot	addition	algorisme	aller
abaisser	1	0.275	0.525	0	0	0	0.025	0.05	0
ajoutement	0.275	1	0.7	0	0.025	0	0	0	0.025
anteriorer	0.525	0.7	1	0.025	0	0	0	0	0
abreger	0	0	0.025	1	0.25	0.325	0	0	0
ajouter	0	0.025	0	0.25	1	0.15	0	0.025	0.125
aliquot	0	0	0	0.325	0.15	1	0	0	0.025
addition	0.025	0	0	0	0	0	1	0.875	0.05
algorisme	0.05	0	0	0	0.025	0	0.875	1	0
aller	0	0.025	0	0	0.125	0.025	0.05	0	1

Table 9: Frequency of neighborhood matrix (same excerpt as 8, with row and columns reorganized)

- Black cells stand for pairs that are neighbors with high probability (proximity happens with frequency greater than $A + B$, here 0.1787).
- White cells stand for pairs that are not neighbors with high probability (proximity happens with frequency less than $A - B$, here 0.0014).
- Grey cells are not conclusive, they are the *fickle pairs*.

If we rearrange the order of cells and columns through Bertin permutations, we immediately make remarkable clustering properties appear (see Table 9.)

For each word, through this treatment we get a list of words that can roughly be grouped around two poles: the strongly associated and the almost never associated ones. Between these two extremes lies a central yet difficult to characterize.

This technique could be used for classification, but here our main objective is a bit different: we are mostly interested in a characterization of words that have high mobility in Kohonen maps, that we call *fickle words*.

2.3. From fickle pairs to fickle words

We call *fickle* a word which belongs to a huge number of fickle pairs:

$$|\{i, |\mathcal{M}_{i,j} - A| \leq B\}| \geq \Theta$$

Unfortunately, it is not quite an easy task to find an appropriate threshold Θ. Here we have decided to fix it according to data interpretation. The 30 ficklest words, whose number of safe neighbors/non-neighbors (non-fickle pairs) is between 89 and 119, are displayed in Figure 10.

contraire "opposite" (89)	*regle de trois* "rule of three" (104)	*depenser* "to expend" (112)
doubler "to double" (89)	*savoir* "to know" (105)	*racine* "root" (113)
falloir "to need" (93)	*partie* "to divide" (105)	*chose* "thing" (113)
meme "same, identical" (93)	*position* "position" (107)	*compter* "to count" (113)
pratique "practical" (94)	*exemple* "for example" (107)	*dire* "to say" (113)
seulement "only" (94)	*demi* "half" (108)	*nombrer* "count" (115)
double "double" (97)	*garder* "to keep"(109)	*raison* "calculation, problem" (116)
multiplication (99)	*science* "science" (109)	*donner* "to give" (117)
reduire "to reduce" (103)	*pouvoir* "can" (111)	*ensemble* "together" (117)
regle "rule" (103)	*se* "if" (111)	*valoir* "to be worth" (119)

Figure 10: 30 ficklest words among 219 studied. For each word, the number between brackets stands for how many non-fickle pairs it belongs to.

2.4. Graph of robust neighborhood

Let us have a different look at the neighborhood matrix (f_{ij}) where f_{ij} is the frequency of two words belonging to the same neighborhood. Instead of trying to jump right ahead and identify fickle words in an absolute way, we can study the robust connections between words *per se*, in order to produce some interesting clustering of the words.

For example, if we have a look at the excerpt from Table 8, we notice immediately that some groups of words are very often in the same neighborhood, while their connections to the rest of the graph are much more hazardous. This initial intuition becomes quite obvious if we reorganize the rows and columns (following Bertin's permutation matrices idea), as we can see on Table 9.

We cannot display here the whole matrix for the 219 forms - in addition, the algorithm for reorganization would not be efficient enough - so we have decided to focus on a specific group of words: the fickle words. Indeed, the fickle words are the most difficult to study, since by definition they do not have a very fixed position on the Kohonen maps, and additionally it appears that they are not well distinguished by Factorial Correspondence Analysis either.

Table 11 shows the frequency matrix for the 30 ficklest words. The clustering is not obvious *a priori*, so we can use a different representation for better visualization of the underlying structures. We can fix the threshold $A+B$ as defined in equation (12) and consider this matrix as the adjacency matrix of a graph $G(V, E)$ such that:

- the set of vertices V is identified to the fickle words
- the set of edges E is defined by $(i, j) \in E \Leftrightarrow f_{ij} > A + B$

In other terms, G is the graph of highly probable neighborhood relations in Kohonen maps. In the case of fickle words, the graph G is given by Figure 12.

2.5. Quasi-cliques

Graphs are powerful tools for visualization, since the graphical representation can be built according to some parameters that ensure highly connected set of vertices to be gathered as much as possible. Still, it can be interesting not to rely only on graphical intuition, but also to use some clustering algorithms with performance guarantee.

Since our graph is pretty dense, it appears that the concept we need here is a quasi-clique coloring. For an introduction to quasi-clique and clique partition problems, one can for example refer to [25]. A *quasi-clique* is a subgraph of highest density; typically, if h is a nondecreasing function, $K \subset V$ is a quasi-clique according to h if $|E[K]| \geq h(|K|, |V|)$. Note that we use the following notations, which are classical in graph theory: if $W \subset V$ is a subset of vertices, then $G[W]$ is the subgraph induced by W, and $E[W]$ is set of edges which are internal to $G[W]$. Here we choose $h : |K|, |V| \mapsto |K|(|K| - 1)/2 - 1$. In other terms, we define a quasi-clique as a subgraph such that every pair of vertices except at most one is connected.

	contraire	doubler	falloir	meme	pratique	seulement	double	multiplication	reduire	regle	regle de trois	partie	savoir	exemple	position	demi	garder	science	pouvoir	se	depenser	chose	compter	dire	racine	nombrer	raison	donner	ensemble	valoir
contraire	1	0	$.6_{25}$	$.7_5$	0	$.0_{25}$	0	$.0_{75}$	.2	$.1_{25}$	$.0_5$	$.0_5$	$.4_{25}$	$.0_5$	$.0_{25}$	$.6_5$	0	0	$.0_5$	$.0_{25}$	$.5_5$	.1	0	0	$.6_{75}$	$.0_{25}$	$.0_5$	$.0_{75}$	$.0_{75}$	$.0_{75}$
doubler	0	1	$.0_5$	$.0_{25}$	$.0_5$	$.0_{75}$	$.9_{25}$	$.0_{75}$	$.0_5$	$.0_5$	$.0_{75}$	$.5_{75}$	$.0_5$	$.0_{25}$	$.1_{75}$	$.0_{25}$	$.2_{25}$	$.3_{25}$	$.0_{75}$	$.2_5$	$.0_{75}$	$.2_{75}$	$.6_5$	.1	$.0_{25}$	.1	$.0_{25}$	$.4_5$	$.3_{75}$	.2
falloir	$.6_{25}$	$.0_5$	1	$.7_{75}$	$.0_{25}$	0	$.0_5$	0	.2	.1	$.0_5$	$.1_{25}$	$.2_5$	$.0_5$	.1	$.5_5$	.1	$.0_5$	.1	$.0_{25}$	$.8_5$	.1	$.0_{75}$	$.1_{25}$	$.6_{75}$	$.0_{25}$	$.1_5$	$.1_{75}$	$.1_5$	$.0_5$
meme	$.7_5$	$.0_{25}$	$.7_{75}$	1	0	$.0_{25}$	$.0_{25}$	$.0_5$	$.2_{75}$	.1	$.0_5$	$.1_{25}$	$.2_{75}$	$.0_5$	$.0_{75}$	$.7_{25}$	$.0_5$	$.0_{25}$	$.0_{75}$	$.0_{25}$	$.7_5$	$.1_{75}$	$.0_5$	$.0_{75}$	$.6_{75}$	$.0_{25}$	$.1_{25}$	$.1_5$	$.1_5$	$.0_5$
pratique	0	$.0_5$	$.0_{25}$	0	1	$.7_{75}$	$.0_5$	$.6_{25}$	$.2_{75}$	$.3_{25}$	.1	$.1_5$	.1	$.3_{75}$	$.0_5$	0	$.0_{75}$	$.0_5$	$.0_5$	$.1_5$	$.0_5$	$.4_{25}$	.1	0	0	$.0_5$	$.0_{25}$	$.1_{25}$	.1	.1
seulement	$.0_{25}$	$.0_{75}$	0	$.0_{25}$	$.7_{75}$	1	$.0_{75}$	$.6_{75}$	$.2_5$	$.2_{75}$	$.1_{25}$	$.1_5$	$.1_{25}$	$.3_5$	$.0_5$	0	$.0_5$	$.0_{75}$	$.0_5$	$.1_{75}$	$.0_{25}$	$.4_{75}$	.1	0	0	$.0_5$	$.0_{25}$	$.1_{75}$	$.0_{75}$	$.0_{25}$
double	0	$.9_{25}$	$.0_5$	$.0_{25}$	$.0_5$	$.0_{75}$	1	$.0_{25}$	$.0_{75}$	$.0_5$	$.1_{75}$	.5	$.0_5$	$.0_5$	$.1_{75}$	$.0_{25}$	$.3_5$	$.2_{75}$	.1	$.2_{75}$	.1	$.2_{75}$	$.6_{25}$	.1	$.0_{25}$	$.0_{75}$	$.0_5$	$.4_{75}$	$.3_{25}$	.2
multiplication	$.0_{75}$	$.0_{75}$	0	$.0_5$	$.6_{25}$	$.6_{75}$	$.0_{25}$	1	.1	$.3_{25}$	.1	$.1_{25}$	$.0_5$	$.0_5$	0	$.0_5$	$.0_{25}$	$.0_{75}$	$.0_{25}$	$.0_{75}$	0	.5	$.0_{25}$	$.0_{25}$	0	$.3_{25}$	$.0_{25}$	$.0_{25}$	$.0_5$	0
reduire	.2	$.0_5$	.2	$.2_{75}$	$.2_{75}$	$.2_5$	$.0_{75}$	.1	1	$.2_{75}$	$.2_{25}$	$.2_{25}$	$.0_{75}$	$.2_{75}$	.1	$.3_{25}$	$.3_{25}$	.1	0	$.1_{25}$	$.1_5$	$.3_{75}$	.1	$.0_5$	$.1_5$	$.0_{75}$	$.0_5$	$.0_{75}$	$.2_{75}$	$.0_5$
regle	$.1_{25}$	$.0_5$	.1	.1	$.3_{25}$	$.2_{75}$	$.0_5$	$.3_{25}$	$.2_{75}$	1	$.0_{25}$	$.2_{25}$	$.1_{75}$	.1	0	$.0_5$	$.0_5$	$.2_5$	$.0_5$	$.0_5$	$.0_{75}$	$.2_{25}$	.1	.3	$.0_5$	$.1_5$	0	$.0_{25}$	$.2_5$	.1
regle de trois	$.0_5$	$.0_{75}$	$.0_5$	$.0_5$	.1	$.1_{25}$	$.1_{75}$	.1	$.2_{25}$	$.0_{25}$	1	$.0_{25}$	0	$.1_5$	.7	$.0_5$	$.6_5$	$.0_{25}$	$.5_5$	$.6_{25}$	$.0_{25}$	$.2_{25}$	$.0_{25}$	$.0_5$	$.0_{25}$	$.0_{75}$	$.6_{75}$	$.4_{75}$	0	.3
partie	$.0_5$	$.5_{75}$	$.1_{25}$	$.1_{25}$	$.1_5$	$.1_5$	.5	$.1_{25}$	$.2_{25}$	$.2_{25}$	$.0_{25}$	1	$.0_{75}$	$.0_5$	.1	$.0_5$	.2	$.5_5$	$.0_5$	$.1_{75}$	$.1_5$	$.3_{75}$	$.6_5$	$.2_{75}$	$.0_{25}$	.1	0	.3	$.7_{75}$	$.1_{25}$
savoir	$.4_{25}$	$.0_5$	$.2_5$	$.2_{75}$	.1	$.1_{25}$	$.0_5$	$.0_5$	$.0_{75}$	$.1_{75}$	0	$.0_{75}$	1	$.1_{25}$	$.0_{75}$	.3	0	$.0_{25}$	.1	$.0_5$	$.1_{75}$	$.0_{75}$	.1	$.0_{25}$	$.2_5$	0	.1	.2	0	.2
exemple	$.0_5$	$.0_{25}$	$.0_5$	$.0_5$	$.3_{75}$	$.3_5$	$.0_5$	$.0_5$	$.2_{75}$	.1	$.1_5$	$.0_5$	$.1_{25}$	1	$.0_{75}$	$.0_{75}$	.1	$.0_5$	$.0_{75}$	.3	$.0_5$	$.1_5$	.1	$.0_{25}$	$.0_{25}$	0	$.1_5$	.3	$.0_{25}$	$.3_{25}$
position	$.0_{25}$	$.1_{75}$	.1	$.0_{75}$	$.0_5$	$.0_5$	$.1_{75}$	0	.1	0	.7	.1	$.0_{75}$	$.0_{75}$	1	$.0_{75}$	.6	$.1_{25}$	$.7_{75}$	$.7_{25}$	$.0_5$	.1	$.0_5$	.2	0	0	$.7_{25}$	$.5_{25}$	$.0_{75}$	$.5_{25}$
demi	$.6_5$	$.0_{25}$	$.5_5$	$.7_{25}$	0	0	$.0_{25}$	$.0_5$	$.3_{25}$	$.0_5$	$.0_5$	$.0_5$	.3	$.0_{75}$	$.0_{75}$	1	$.0_{75}$	0	$.0_{75}$	$.0_5$	$.5_{75}$	$.0_{75}$	0	$.0_{25}$	$.6_{25}$	$.0_{25}$	$.1_{75}$	$.1_5$	$.0_{75}$	$.0_5$
garder	0	$.2_{25}$	.1	$.0_5$	$.0_{75}$	$.0_5$	$.3_5$	$.0_{25}$	$.3_{25}$	$.0_5$	$.6_5$	.2	0	.1	.6	$.0_{75}$	1	.2	$.4_{25}$	.5	$.0_{75}$	.2	$.1_{75}$	.1	$.0_{25}$	$.0_{75}$	$.4_{75}$	$.3_{25}$	$.1_{75}$	$.1_{25}$
science	0	$.3_{25}$	$.0_5$	$.0_{25}$	$.0_5$	$.0_{75}$	$.2_{75}$	$.0_{75}$	.1	$.2_5$	$.0_{25}$	$.5_5$	$.0_{25}$	$.0_5$	$.1_{25}$	0	.2	1	$.1_{25}$	.1	$.0_{25}$	$.1_5$	$.4_{75}$	$.3_5$	0	.1	$.0_{25}$	$.0_{75}$	$.6_{25}$	$.0_{75}$
pouvoir	$.0_5$	$.0_{75}$	.1	$.0_{75}$	$.0_5$	$.0_5$	.1	$.0_{25}$	0	$.0_5$	$.5_5$	$.0_5$	.1	$.0_{75}$	$.7_{75}$	$.0_{75}$	$.4_{25}$	$.1_{25}$	1	$.6_5$	$.0_{75}$	$.0_{25}$	$.0_{25}$	$.2_{25}$	$.0_{25}$	$.0_{25}$	.8	$.3_{25}$	$.0_5$	$.4_5$
se	$.0_{25}$	$.2_5$	$.0_{25}$	$.0_{25}$	$.1_5$	$.1_{75}$	$.2_{75}$	$.0_{75}$	$.1_{25}$	$.0_5$	$.6_{25}$	$.1_{75}$	$.0_5$	.3	$.7_{25}$	$.0_5$	.5	.1	$.6_5$	1	0	.2	$.1_5$	$.0_{75}$	0	$.0_5$	$.6_5$	$.6_{75}$	$.1_{25}$	$.5_5$
depenser	$.5_5$	$.0_{75}$	$.8_5$	$.7_5$	$.0_5$	$.0_{25}$	.1	0	$.1_5$	$.0_{75}$	$.0_{25}$	$.1_5$	$.1_{75}$	$.0_5$	$.0_5$	$.5_{75}$	$.0_{75}$	$.0_{25}$	$.0_{75}$	0	1	$.1_{75}$	$.1_{75}$	$.0_{25}$	.8	$.0_{25}$	$.0_{75}$	$.1_{75}$	.1	0
chose	.1	$.2_{75}$	.1	$.1_{75}$	$.4_{25}$	$.4_{75}$	$.2_{75}$	.5	$.3_{75}$	$.2_{25}$	$.2_{25}$	$.3_{75}$	$.0_{75}$	$.1_5$	.1	$.0_{75}$	.2	$.1_5$	$.0_{25}$	.2	$.1_{75}$	1	$.4_{25}$	$.0_{25}$	$.0_{75}$	$.2_5$	$.0_{25}$	$.1_5$	$.1_5$	$.0_5$
compter	0	$.6_5$	$.0_{75}$	$.0_5$	.1	.1	$.6_{25}$	$.0_{25}$	.1	.1	$.0_{25}$	$.6_5$	.1	.1	$.0_5$	0	$.1_{75}$	$.4_{75}$	$.0_{25}$	$.1_5$	$.1_{75}$	$.4_{25}$	1	$.0_{75}$	0	$.0_5$	0	$.3_{75}$	$.4_{25}$	$.1_{25}$
dire	0	.1	$.1_{25}$	$.0_{75}$	0	0	.1	$.0_{25}$	$.0_5$	.3	$.0_5$	$.2_{75}$	$.0_{25}$	$.0_{25}$	.2	$.0_{25}$	.1	$.3_5$	$.2_{25}$	$.0_{75}$	$.0_{25}$	$.0_{25}$	$.0_{75}$	1	0	$.0_{25}$	$.0_5$	$.0_{25}$	$.4_{25}$	$.1_{25}$
racine	$.6_{75}$	$.0_{25}$	$.6_{75}$	$.6_{75}$	0	0	$.0_{25}$	0	$.1_5$	$.0_5$	$.0_{25}$	$.0_{25}$	$.2_5$	$.0_{25}$	0	$.6_{25}$	$.0_{25}$	0	$.0_{25}$	0	.8	$.0_{75}$	0	0	1	$.0_{75}$	0	$.0_{25}$	$.0_5$	$.0_{25}$
nombrer	$.0_{25}$	.1	$.0_{25}$	$.0_{25}$	$.0_5$	$.0_5$	$.0_{75}$	$.3_{25}$	$.0_{75}$	$.1_5$	$.0_{75}$	.1	0	0	0	$.0_{25}$	$.0_{75}$	.1	$.0_{25}$	$.0_5$	$.0_{25}$	$.2_5$	$.0_5$	$.0_{25}$	$.0_{75}$	1	0	0	.1	0
raison	$.0_5$	$.0_{25}$	$.1_5$	$.1_{25}$	$.0_{25}$	$.0_{25}$	$.0_5$	$.0_{25}$	$.0_5$	0	$.6_{75}$	0	.1	$.1_5$	$.7_{25}$	$.1_{75}$	$.4_{75}$	$.0_{25}$	.8	$.6_5$	$.0_{75}$	$.0_{25}$	0	$.0_5$	0	0	1	$.3_{75}$	0	$.4_{25}$
donner	$.0_{75}$	$.4_5$	$.1_{75}$	$.1_5$	$.1_{25}$	$.1_{75}$	$.4_{75}$	$.0_{25}$	$.0_{75}$	$.0_{25}$	$.4_{75}$	.3	.2	.3	$.5_{25}$	$.1_5$	$.3_{25}$	$.0_{75}$	$.3_{25}$	$.6_{75}$	$.1_{75}$	$.1_5$	$.3_{75}$	$.0_{25}$	$.0_{25}$	0	$.3_{75}$	1	$.1_{75}$	$.5_5$
ensemble	$.0_{75}$	$.3_{75}$	$.1_5$	$.1_5$	.1	$.0_{75}$	$.3_{25}$	$.0_5$	$.2_{75}$	$.2_5$	0	$.7_{75}$	0	$.0_{25}$	$.0_{75}$	$.0_{75}$	$.1_{75}$	$.6_{25}$	$.0_5$	$.1_{25}$	.1	$.1_5$	$.4_{25}$	$.4_{25}$	$.0_5$	.1	0	$.1_{75}$	1	$.1_{25}$
valoir	$.0_{75}$	.2	$.0_5$	$.0_5$	.1	$.0_{25}$	.2	0	$.0_5$	.1	.3	$.1_{25}$	.2	$.3_{25}$	$.5_{25}$	$.0_5$	$.1_{25}$	$.0_{75}$	$.4_5$	$.5_5$	0	$.0_5$	$.1_{25}$	$.1_{25}$	$.0_{25}$	0	$.4_{25}$	$.5_5$	$.1_{25}$	1

Table 11: Frequency of neighborhood matrix for the ficklest words only = adjacency matrix of the neighborhood graph of the ficklest

Unfortunately, finding a maximum quasi-clique is NP-hard in the general case as well as with this specific function. Still since the graph is small and has bounded degree, we can afford to use moderately exponential algorithms (see [26]). We build a partition of the graph as such:

Algorithm Glutton Quasi-Clique Decomposition(G)

1: **if** $\kappa = \{H \subset V,\ |H| \geq 4 \vee E[H] \leq |H|(|H|-1)/2-1)\} \neq \emptyset$ **then**
2: Find K which is maximum among κ
3: Return $(K, \text{GLUTTON QUASI} - \text{CLIQUE DECOMPOSITION}(G[V \setminus K]))$
4: **else**
5: Return V

This formal definition can be rephrased with a simple explanation: the algorithm will look for the maximum size quasi-clique, add it as an item of the partition and proceed recursively until the graph contains no quasi-clique of size 4 or more. The remaining vertices are left isolated in the decomposition (note that a quasi-clique of size 3 is simply a path, and thus not very interesting to study).

Of course the difficult point is the computation of κ at each step. There are basically two solutions, depending on the density of the graph.

If the graph $G(V, E)$ has high average degree $\delta_G = 2|E|/|V|$, then its complementary graph $\bar{G}(V, V^2 \setminus E)$ has low average degree $\delta_{\bar{G}} = |V| - 1 - \delta_G$, and thus the algorithm from [26] is efficient.

On the other hand, if the graph has low average degree, we can use the following algorithm,

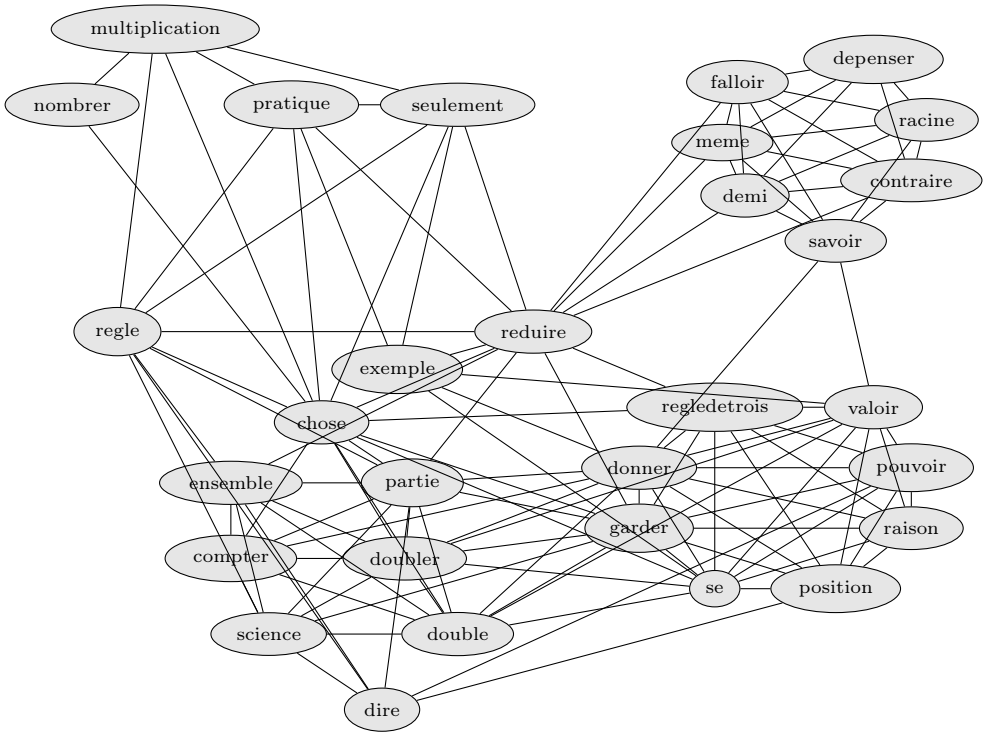

Figure 12: Graph of the relations between fickle words. Two nodes are connected if the words are significantly neighbors.

that basically solves the quasi-clique problem through the resolution of a small (quadratic) number of clique problems:

Algorithm QUASI CLIQUE(V, E)

1: $K = \text{CLIQUE}(V, E)$
2: **for all** $u, v \in V (u, v) \notin E$ **do**
3: $\quad K = \max\{K, \text{CLIQUE}(V, E \cup (u, v))\}$
4: **return** K

Here CLIQUE can be any exact algorithm for the maximum clique problem, which is **NP**-hard too. To our knowledge, the fastest ones are those designed in [27].

3. Analysis of results

3.1. Robust Kohonen maps

In what follows, we consider the Kohonen map after removing the fickle words (in gray), see Figure 13. We call this modified map a Robust Kohonen map.

The Robust Kohonen map shows a contrast between the top right corner and the bottom left one, the same contrast as between left and right sides on the first axis in FCA representation (see Figure 2).

Indeed, the top right corner contains words linked to arithmetic practice and verbs that are used to build arithmetical operations as *retenir*"to retain", *emprunter*"to borrow", etc. No manuscript is specific of this part of the map, even if the BNF fr. 2050 and *Kadran aux Marchans* appear in the low part of the graph.

On the bottom left corner, one finds the lexical inheritance of the medieval university represented by the manuscript BNF fr. 1339. Some texts contain a highly specialized vocabulary, with connections to the university world, with words such as *article* "article", *algorism* "algorithm",

sain "integer", or a vocabulary of geometry used for the extraction of roots (*carree*).

The other two corners of the map are characterized by vocabulary taken from two specific manuscripts. The BSG 3143 in the up left corner is a treatise written by Jean Adam for future Louis XI, that is exceptional in the corpus because it uses Latin words and roman numbers and also because it had to be pleasant for the prince. In spite of this, it shares with the Nantes 456 and BNF fr.1339 words as *gecter*, *gectons* that are marks of abacus algorithms.

In the opposite corner, the *Traicté de la pratique* is marked by a more descriptive vocabulary of mathematical problems (*item* "item", *demande* "demand", *requerir* "to call for", *quant* "quant") and stronger scientific approach (*aliquot* "aliquot", *corps* "field", *proportionnellement* "proportionally".)

What can we do with the list of "fickle words" from this map ? First, it is remarkable that a part of fickle words concerns the algorithm of the rule of three. This algorithm consists of a "multiplication" (*multiplier*) by the "opposite" (*contraire*) and of a "division" (*diviser*). Other fickle words are related to the operations (*reduction* "fractions reducing", *multiplier* "to multiply", *additionner* "to add"), and with words having a distinctive didactic flavor (*falloir* "have to", *dire* "to say"). As a matter of fact, the two main technical issues for these XV$^{\text{th}}$ century authors are to teach how to use the rule of three and the fractions to their readers.

3.2. Improved visualization for FCA

The combination of both techniques FCA and SOM whose result is displayed in Figures 14 and 15 is interesting because it preserves properties of the FCA while giving additional information about the center of the projection - which is usually difficult to interpret. Indeed, the identification of the fickle words on the FCA projections allows us to improve the general interpretation of the factorial graphs, where some words are located because of the algorithm and not because of their attraction to other words and to the texts.

Remember that, on the first two factorial axes (see Figure 14), we have observed an opposition between the university legacy, on the right, and a more practical pole with rules, problems and fractions, on the left. It could be tempting to support this observation with very significant words such as *pratique* "practical" or *règle de trois* "rule of three". Still, the enhancement of the fickle forms on the FCA shows that these words are in fact shared between many different texts and not only linked to the more 'practical' ones: Nicolas Chuquet and *Traicté en la praticque*. As a matter of fact, they do belong to all the texts.

It is the same for two other words (*raison* "reason", *dire* "to say"). The word *raison* is an ambiguous word, in a way, because it can mean calculation, with textual matches like "do your reasons", or indicate mathematical problems. "To say" ranks sixth by order of frequency among verbs in the corpus. Note that all most important verbs are not fickle words. The first eight verbs for occurrence are: *être*(14523) "to be", *avoir*(4563) "to have", *devoir*(3826) "must", *faire*(3431) "to do", *multiplier*(3228) "to multiply", *dire*(2461) "to say", *partir*(2648) "to divide", "valoir"(2606) "to be worth". One of the particular meanings of "to say" comes from the orality of this type of text. Understanding arithmetic operations often supposes saying it aloud.

Another interest of this kind of representation is the interpretation of the center of FCA. As we can see in Figure 16, fickle words are close to the center, but there are other words in the same place, which we could interpret.

To conclude, we can see that two levels of interpretation are superimposed: the fickle pairs reveal the shared lexicon and the factorial map inserts them in a local interaction system. And since the fickle words list is computed independently from the FCA, we can successively study these interactions on each axis. It is the articulation between these two levels which makes this representation interesting. At the end, the meaning of this new kind of factorial map is quite intuitive and offers easy tools to the argumentation.

minutes super*		notes	calculer cubic	fois mettre **BNF10259**		contraire depenser falloir meme racine	aller donc ensuivre savoir	multiplier	regle venir
gecter					dessous	barrater demi	somme	voir	assembler
	BSG3143	parteur	defaillir duplation mediation nommer numeration senestre	semblables		circulaires demeurer derenier disaine ecrire entendre formes nombrateur oter prouver	entrer laisser rien	emprunter figure regarder	figure de non rien fraction muer rayes retenir
notables		denominations multiplicateur nominateur ordonne	abaisser comptes endroit proposer	anteriorer diminution enseignement enseigner moyen signifiant surplus trancher	possible reduire	difficile progression repondre	avaluer	faillir	monter partiteur
bref gectons multiplication pratique seulement	generale latin nombrer proportion reduction unite	soustraction	denominateur	entier	ajoutement		remotion **Kadran**		**BNF2050**
chose	nulle	ensemble partie	**Nantes456**				partement valoir		etre
arithmetique compter preuve tenir	cubbement destre digit diviser diviseur division lignes nombre compo sequerir	dire figurer poser science			abreger lever precedent quotiens	numerateur	moindre nombre **Nicolas Chuquet**	partir plus raison	
egalir egaliser especes question total traiter	mesurer	grand	apparaitre position soustraire	ajouter	leurs prendre quant quantefois reponse trouver		part	commun Item	devoir droit exemple reste rester
algorisme article carrees cercle envient ligne pair sain **BNF1339**	addition chiffre former	double doubler moitie	appeler donner maniere pouvoir se	bailler demander mises nomper pareillement vouloir		egale faire montrer necessaire romp selon			
cautelle	demontrer dessus	garder regle de trois	aliquot composer corps moins sub* toutefois	partant plaisir proportionellement requerir residu	appartenir convenir demande difference egaulx maieur millions rate survendre tant		naturel roupt **Traicte praticque**		fausse

Figure 13: Robust Kohonen map: fickle words are removed (in gray)

3.3. Neighborhood graphs

Bertin matrix (see Table 17) shows some remarkable clustering among fickle words. The question is now to produce some interpretation of this clustering.

First, we can observe that the Bertin matrix displays four groups along its diagonal, from the more connected on the top left, to the less connected on the bottom right.

The first list (*contraire*, *depenser*, *falloir*, *racine*, *meme*, *demi*, *savoir*, see Figure 10 for a translation) is a collection of rather heterogeneous words. There are words frequently used such as *demi* and others bearing a strong polysemy such as *falloir*.

A possible explanation may be that these groups of words form phrases in the corpus (that is usually called *textual co-occurence): the words are spaced only one or two words from each other.* and that these associations are reflected in the table. In that case, fickle words can be used as a

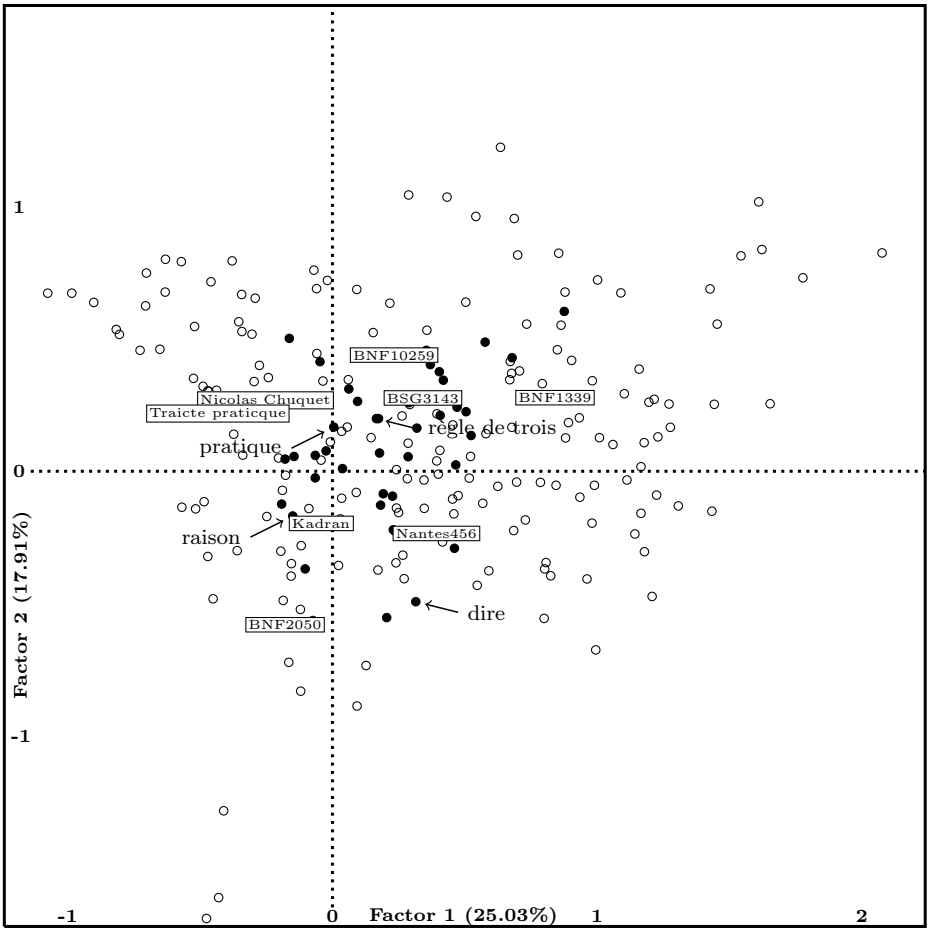

Figure 14: Projection on first two factors of the FCA. Only the fickle words are in black.

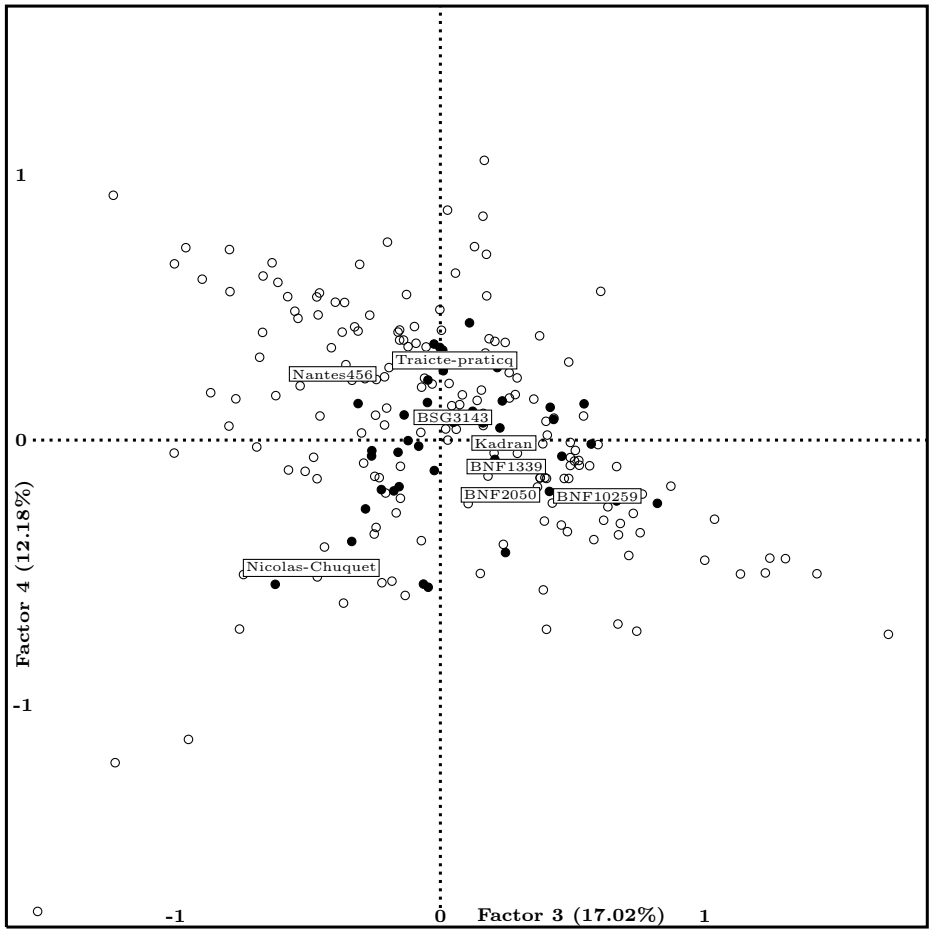

Figure 15: Projection on third and fourth factors of the FCA. Only the fickle words are in black.

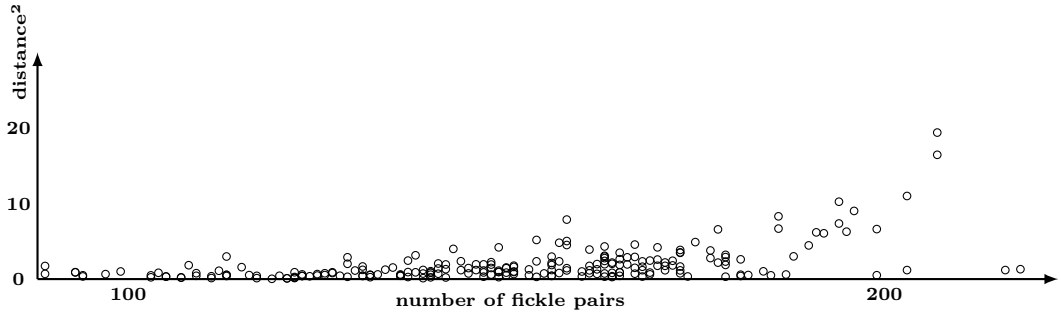

Figure 16: Correlation between fickleness and distance to the center. x-axis represents the number of fickle pairs a word belongs to, while y-axis stands for the square distance to the origin.

tool to extract *topoi*. Thus, for example, *savoir* "to know" and *contraire* "contrary" are often used in phrases such as *savoir par son contraire* "to know smth through its contrary".

We can also think that these clustering properties reveal more distant co-occurrences, that means words appearing in the same sentence or paragraph, but not necessarily the same *topos*. For instance, *falloir* "to have to do" has a lot of such co-occurrences with fickle words like *reduire* "to reduce", *racine* "root" and *savoir* "to know". In this configuration, we can in fact suppose that all these words are shared by the same sentences.

	contraire	depenser	falloir	racine	meme	demi	savoir	valoir	regledetrois	raison	position	pouvoir	garder	donner	se	doubler	double	partie	science	compter	ensemble	dire	regle	chose	reduire	seulement	pratique	multiplication	exemple	nombrer
contraire	1	.5$_5$	.6$_{25}$	.6$_{75}$	.7$_5$	.6$_5$	.4$_{25}$	.0$_{75}$	.0$_5$	.0$_5$	.0$_{25}$	.0$_5$	0	.0$_{75}$	.0$_{25}$	0	0	.0$_5$	0	0	.0$_{75}$	0	.1$_{25}$	.1	.2	.0$_{25}$	0	.0$_{75}$	.0$_5$	.0$_{25}$
depenser	.5$_5$	1	.8$_5$	.8	.7$_5$	.5$_{75}$	.1$_{75}$	0	.0$_{25}$	.0$_{75}$	.0$_5$	.0$_{75}$	.0$_{75}$	.1$_{75}$	0	.0$_{75}$	.1	.1$_5$	.0$_{25}$	.1$_{75}$	.1	.0$_{25}$	.0$_{75}$	.1$_{75}$	.1$_5$	.0$_{25}$	.0$_5$	0	.0$_5$	.0$_{25}$
falloir	.6$_{25}$	.8$_5$	1	.6$_{75}$	.7$_{75}$	.5$_5$	.2$_5$	.0$_5$	.0$_5$	.1$_5$	.1	.1	.1	.1$_{75}$	.0$_{25}$	.0$_5$	.0$_5$	.1$_{25}$	.0$_5$	.0$_{75}$	.1$_5$	.1$_{25}$	.1	.1	.2	0	.0$_{25}$	0	.0$_5$	.0$_{25}$
racine	.6$_{75}$	.8	.6$_{75}$	1	.6$_{75}$	.6$_{25}$	.2$_5$	.0$_{25}$	.0$_{25}$	0	0	.0$_{25}$	.0$_{25}$	.0$_{25}$	0	.0$_{25}$	.0$_{25}$	.0$_{25}$	0	0	.0$_5$	0	.0$_5$	.0$_{75}$	.1$_5$	0	0	0	.0$_{25}$	.0$_{75}$
meme	.7$_5$	.7$_5$	.7$_{75}$	.6$_{75}$	1	.7$_{25}$	.2$_{75}$	.0$_5$	.0$_5$	.1$_{25}$	.0$_{75}$	.0$_{75}$	.0$_5$	.1$_5$	.0$_{25}$	.0$_{25}$	.0$_{25}$	.1$_{25}$	.0$_{25}$	.0$_5$	.1$_5$	.0$_{75}$	.1	.1$_{75}$	.2$_{75}$	.0$_{25}$	0	.0$_5$	.0$_5$	.0$_{25}$
demi	.6$_5$	.5$_{75}$	.5$_5$	.6$_{25}$	.7$_{25}$	1	.3	.0$_5$	.0$_5$	.1$_{75}$	.0$_{75}$	.0$_{75}$	.0$_{75}$	.1$_5$	.0$_5$	.0$_{25}$	.0$_{25}$	.0$_5$	0	0	.0$_{75}$	.0$_{25}$	.0$_5$	.0$_{75}$	.3$_{25}$	0	0	.0$_5$	.0$_{75}$	.0$_{25}$
savoir	.4$_{25}$	.1$_{75}$	.2$_5$	.2$_5$	.2$_{75}$	.3	1	.2	0	.1	.0$_{75}$	.1	0	.2	.0$_5$	.0$_5$	.0$_5$	.0$_{75}$	.0$_{25}$	.1	0	.0$_{25}$	.1$_{75}$	.0$_{75}$	.0$_{75}$	.1$_{25}$	.1	.0$_5$	.1$_{25}$	0
valoir	.0$_{75}$	0	.0$_5$	.0$_{25}$	.0$_5$	.0$_5$	.2	1	.3	.4$_{25}$	.5$_{25}$	.4$_5$	.1$_{25}$	.5$_5$	.5$_5$	.2	.2	.1$_{25}$	.0$_{75}$	.1$_{25}$	.1$_{25}$	.1$_{25}$	.1	.0$_5$	.0$_5$	.0$_{25}$	.1	0	.3$_{25}$	0
regledetrois	.0$_5$	.0$_{25}$	.0$_5$	.0$_{25}$	.0$_5$	.0$_5$	0	.3	1	.6$_5$	.7	.5$_5$	.6$_5$	.4$_{75}$	.6$_{25}$	.0$_{75}$	.1$_{75}$	.0$_{25}$	.0$_{25}$	.0$_{25}$	0	.0$_5$	.0$_{25}$	.2$_{25}$	.2$_{25}$	.1$_{25}$	.1	.1	.1$_5$	.0$_{75}$
raison	.0$_5$	.0$_{75}$	.1$_5$	0	.1$_{25}$	.1$_{75}$	.1	.4$_{25}$	.6$_{75}$	1	.7$_{25}$	.8	.4$_{75}$	.3$_{75}$	.6$_5$	.0$_{25}$	.0$_5$	0	.0$_{25}$	0	0	.0$_5$	0	.0$_{25}$	.0$_5$	.0$_{25}$	.0$_{25}$	.0$_{25}$	.1$_5$	0
position	.0$_{25}$	.0$_5$	.1	0	.0$_{75}$	.0$_{75}$	.0$_{75}$	.5$_{25}$	.7	.7$_{25}$	1	.7$_{75}$	.6	.5$_{25}$	.7$_{25}$	.1$_{75}$	.1$_{75}$	.1	.1$_{25}$	.0$_5$	.0$_{75}$	.2	0	.1	.1	.0$_5$	.0$_5$	0	.0$_{75}$	0
pouvoir	.0$_5$	.0$_{75}$	.1	.0$_{25}$	.0$_{75}$	.0$_{75}$	.1	.4$_5$	.5$_5$	.8	.7$_{75}$	1	.4$_{25}$	.3$_{25}$	.6$_5$	.0$_{75}$	.1	.0$_5$	.1$_{25}$	.0$_{25}$	.0$_5$	.2$_{25}$	.0$_5$	.0$_{25}$	0	.0$_5$	.0$_5$	.0$_{25}$	.0$_{75}$	.0$_{25}$
garder	0	.0$_{75}$	.1	.0$_{25}$	.0$_5$	.0$_{75}$	0	.1$_{25}$	.6$_5$	.4$_{75}$	.6	.4$_{25}$	1	.3$_{25}$	.5	.2$_{25}$	.3$_5$	.2	.2	.1$_{75}$	.1$_{75}$	.1	.0$_5$	.2	.3$_{25}$	.0$_5$	.0$_{75}$	.0$_{25}$	.1	.0$_{75}$
donner	.0$_{75}$	.1$_{75}$	.1$_{75}$	.0$_{25}$	.1$_5$	.1$_5$	.2	.5$_5$	.4$_{75}$	.3$_{75}$	.5$_{25}$	.3$_{25}$	.3$_{25}$	1	.6$_{75}$	.4$_5$	.4$_{75}$	.3	.0$_{75}$	.3$_{75}$	.1$_{75}$	.0$_{25}$	.0$_{25}$	.1$_5$	.0$_{75}$	.1$_{75}$	.1$_{25}$	.0$_{25}$	.3	0
se	.0$_{25}$	0	.0$_{25}$	0	.0$_{25}$	.0$_5$	.0$_5$	.5$_5$	.6$_{25}$	.6$_5$	.7$_{25}$	.6$_5$	.5	.6$_{75}$	1	.2$_5$	.2$_{75}$	.1$_{75}$	.1	.1$_5$	.1$_{25}$	.0$_{75}$	.0$_5$	.2	.1$_{25}$	.1$_{75}$	.1$_5$	.0$_{75}$	.3	.0$_5$
doubler	0	.0$_{75}$	.0$_5$	.0$_{25}$	.0$_{25}$	.0$_{25}$	.0$_5$	.2	.0$_{75}$	.0$_{25}$	.1$_{75}$	.0$_{75}$	.2$_{25}$	.4$_5$	.2$_5$	1	.9$_{25}$	.5$_{75}$	.3$_{25}$	.6$_5$	.3$_{75}$	.1	.0$_5$	.2$_{75}$	.0$_5$	.0$_{75}$	.0$_5$	.0$_{75}$	.0$_{25}$	.1
double	0	.1	.0$_5$	.0$_{25}$	.0$_{25}$	.0$_{25}$	.0$_5$	.2	.1$_{75}$	.0$_5$	.1$_{75}$	.1	.3$_5$	.4$_{75}$	.2$_{75}$	.9$_{25}$	1	.5	.2$_{75}$	.6$_{25}$	.3$_{25}$	.1	.0$_5$	.2$_{75}$	.0$_{75}$	.0$_{75}$	.0$_5$	.0$_{25}$	.0$_5$	.0$_{75}$
partie	.0$_5$	.1$_5$	.1$_{25}$	.0$_{25}$	.1$_{25}$	.0$_5$	.0$_{75}$	.1$_{25}$	.0$_{25}$	0	.1	.0$_5$	.2	.3	.1$_{75}$	.5$_{75}$	.5	1	.5$_5$	.6$_5$	.7$_{75}$	.2$_{75}$	.2$_{25}$	.3$_{75}$	.2$_{25}$	.1$_5$	.1$_5$	.1$_{25}$	.0$_5$	.1
science	0	.0$_{25}$	.0$_5$	0	.0$_{25}$	0	.0$_{25}$	.0$_{75}$	.0$_{25}$	.0$_{25}$	.1$_{25}$	.1$_{25}$	.2	.0$_{75}$	.1	.3$_{25}$	.2$_{75}$	.5$_5$	1	.4$_{75}$	.6$_{25}$	.3$_5$	.2$_5$	.1$_5$	.1	.0$_{75}$	.0$_5$	.0$_{75}$	.0$_5$	.1
compter	0	.1$_{75}$	.0$_{75}$	0	.0$_5$	0	.1	.1$_{25}$	.0$_{25}$	0	.0$_5$	.0$_{25}$	.1$_{75}$	.3$_{75}$	.1$_5$	.6$_5$	.6$_{25}$	.6$_5$	.4$_{75}$	1	.4$_{25}$	.0$_{75}$	.1	.4$_{25}$	.1	.1	.1	.0$_{25}$	.1	.0$_5$
ensemble	.0$_{75}$	.1	.1$_5$	.0$_5$	.1$_5$	.0$_{75}$	0	.1$_{25}$	0	0	.0$_{75}$	.0$_5$	.1$_{75}$	.1$_{75}$	.1$_{25}$	.3$_{75}$	.3$_{25}$	.7$_{75}$	.6$_{25}$	.4$_{25}$	1	.4$_{25}$	.2$_5$	.1$_5$	.2$_{75}$	.0$_{75}$	.1	.0$_5$	.0$_{25}$	.1
dire	0	.0$_{25}$	.1$_{25}$	0	.0$_{75}$	.0$_{25}$	.0$_{25}$	.1$_{25}$	.0$_5$	.0$_5$	.2	.2$_{25}$	.1	.0$_{25}$	.0$_{75}$	.1	.1	.2$_{75}$	.3$_5$	.0$_{75}$	.4$_{25}$	1	.3	.0$_{25}$	.0$_5$	0	0	.0$_{25}$	.0$_{25}$	.0$_{25}$
regle	.1$_{25}$	.0$_{75}$	.1	.0$_5$	.1	.0$_5$	.1$_{75}$	.1	.0$_{25}$	0	0	.0$_5$	.0$_5$	.0$_{25}$	.0$_5$	.0$_5$	.0$_5$	.2$_{25}$	.2$_5$	.1	.2$_5$	.3	1	.2$_{25}$	.2$_{75}$	.2$_{75}$	.3$_{25}$	.3$_{25}$	.1	.1$_5$
chose	.1	.1$_{75}$	.1	.0$_{75}$	.1$_{75}$	.0$_{75}$	.0$_{75}$	.0$_5$	.2$_{25}$	.0$_{25}$	.1	.0$_{25}$	.2	.1$_5$	.2	.2$_{75}$	.2$_{75}$	.3$_{75}$	.1$_5$	.4$_{25}$	.1$_5$	.0$_{25}$	.2$_{25}$	1	.3$_{75}$	.4$_{75}$	.4$_{25}$	.5	.1$_5$	.2$_5$
reduire	.2	.1$_5$	.2	.1$_5$	.2$_{75}$	.3$_{25}$	.0$_{75}$	.0$_5$	.2$_{25}$	.0$_5$	.1	0	.3$_{25}$	.0$_{75}$	.1$_{25}$	.0$_5$	.0$_{75}$	.2$_{25}$	.1	.1	.2$_{75}$	.0$_5$	.2$_{75}$	.3$_{75}$	1	.2$_5$	.2$_{75}$	.1	.2$_{75}$	.0$_{75}$
seulement	.0$_{25}$	.0$_{25}$	0	0	.0$_{25}$	0	.1$_{25}$	.0$_{25}$	.1$_{25}$	.0$_{25}$	.0$_5$	.0$_5$	.0$_5$	.1$_{75}$	.1$_{75}$	.0$_{75}$	.0$_{75}$	.1$_5$	.0$_{75}$	.1	.0$_{75}$	0	.2$_{75}$	.4$_{75}$	.2$_5$	1	.7$_5$	.6$_{75}$	.3$_5$	.0$_5$
pratique	0	.0$_5$	.0$_{25}$	0	0	0	.1	.1	.1	.0$_{25}$	.0$_5$	.0$_5$	.0$_{75}$	.1$_{25}$	.1$_5$	.0$_5$	.0$_5$	.1$_5$	.0$_5$	.1	.1	0	.3$_{25}$	.4$_{25}$	.2$_{75}$	.7$_5$	1	.6$_{25}$	.3$_{75}$	.0$_5$
multiplication	.0$_{75}$	0	0	0	.0$_5$	.0$_5$	.0$_5$	0	.1	.0$_{25}$	0	.0$_{25}$	.0$_{25}$	.0$_{25}$	.0$_{75}$	.0$_{75}$	.0$_{25}$	.1$_{25}$	.0$_{75}$	.0$_{25}$	.0$_5$	.0$_{25}$	.3$_{25}$	.5	.1	.6$_{75}$	.6$_{25}$	1	.0$_5$	.3$_{25}$
exemple	.0$_5$	.0$_5$	.0$_5$	.0$_{25}$	.0$_5$	.0$_{75}$	.1$_{25}$	.3$_{25}$	.1$_5$	.1$_5$	.0$_{75}$	.0$_{75}$	.1	.3	.3	.0$_{25}$	.0$_5$	.0$_5$	.0$_5$	.1	.0$_{25}$	.0$_{25}$	.1	.1$_5$	.2$_{75}$	.3$_5$	.3$_{75}$	.0$_5$	1	0
nombrer	.0$_{25}$	.0$_{25}$	.0$_{25}$	.0$_{75}$	.0$_{25}$	.0$_{25}$	0	0	.0$_{75}$	0	0	.0$_{25}$	.0$_{75}$	0	.0$_5$	.1	.0$_{75}$	.1	.1	.0$_5$	.1	.0$_{25}$	.1$_5$	.2$_5$	.0$_{75}$	.0$_5$	.0$_5$	.3$_{25}$	0	1

Table 17: Same as Table 11, reorganized and with shades proportional to value.

On the contrary, *demi*, that is a part of the same well-connected group (according to the clustering), does not have any specific co-occurrences in the texts with any word in this group. That is especially interesting, since it reveals the existence of connections that could not have been deduced from a simple study of co-occurrence with classical tools.

The Figure 18 opens another perspective. Indeed, it shows the words that make the link between clusters. These fickle words have a lot of different affinities. We can see, for example, that the positions of *reduire* "to reduce" and *exemple* "example" are not very surprising, because these words are used a lot, in every text, in sentences associating them with various other fickle words, such as "in all the examples preceding the problems", or "the problem of reducing or converting the monetary values".

These questions are not solved yet, and the answer cannot be sure without an enlargement of the corpus. Indeed, we would like to test this hypothesis by using the process described here on a larger part of the corpus.

Conclusion

In this work, we have shown how to use the Kohonen maps as a complement of Factorial Correspondence Analysis methods (FCA)classically used in lexicometry,

- to improve the information provided by the different projections of the FCA,
- to make the Kohonen maps more robust with respect to the randomness of the SOM algorithm, by distinguishing stable neighbor pairs from fickle pairs,
- to build graphs of connections between fickle words which are difficult to analyze by both FCA and Kohonen map alone.

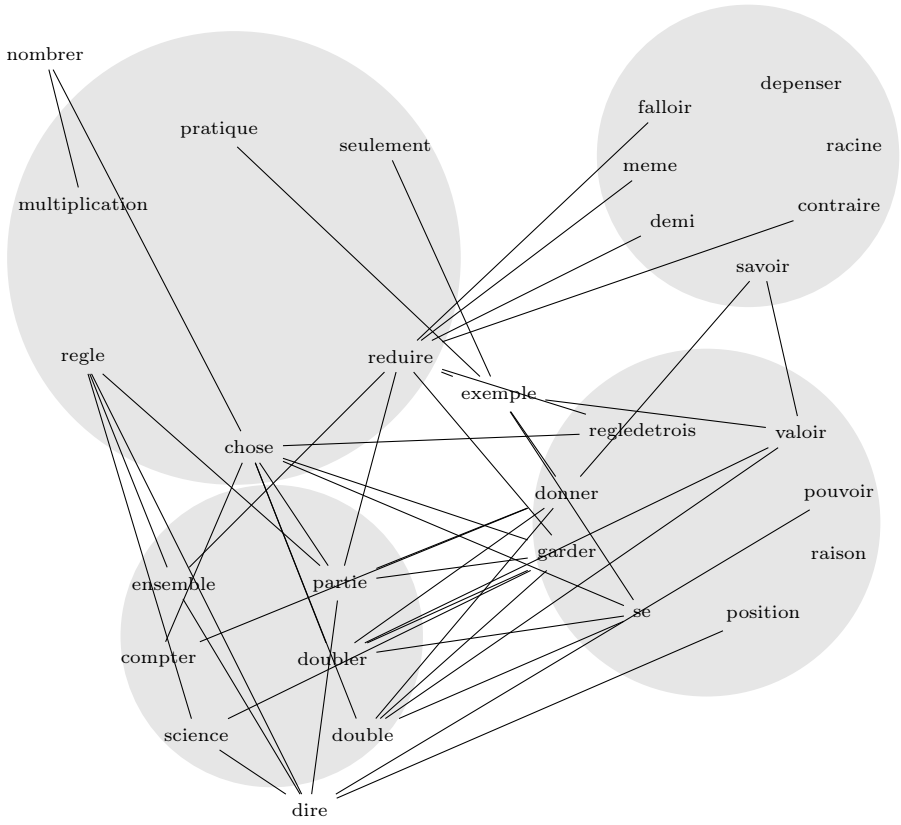

Figure 18: Glutton decomposition in quasi-cliques of maximum size.

We think that it will be interesting to use this methodology on a large variety of corpus, such as political speeches, chivalric culture [28] texts and scientific articles.

[1] N. Bourgeois, M. Cottrell, B. Deruelle, S. Lamassé, P. Letrémy, Lexical recount between factor analysis and kohonen map: Mathematical vocabulary of arithmetic in the vernacular language of the late middle ages, in: P. E. et al. (Ed.), Advances in Self-Organizing Maps, WSOM 2012, Vol. 198, Springer-Verlag Berlin Heidelberg, Santiago, Chili, 2012, pp. 255–264.

[2] S. Lamassé, Les traités d'arithmétique médiévale et la constitution d'une langue de spécialité, in: J. Ducos (Ed.), Sciences et langues au Moyen Âge, Actes de l'Atelier franco-allemand, Paris, 27-30 janvier 2009, Universitätesverlag, Heidelberg, 2012, pp. 66–104.

[3] G. Beaujouan, The place of Nicolas Chuquet in a typology of fifteenth-century french arithmetics, Clarendon Press, 1988, pp. 73–88.

[4] M. Spiesser, Une arithmétique commerciale du XVe siècle : le Compendy de la praticque des nombres de Barthélemy de Romans, De diversis artibus ; t. 70, Brepols, 2003.

[5] P. Benoit, Recherches sur le vocabulaire des opérations élémentaires dans les arithmétiques en langue française de la fin du moyen age, Documents pour l'histoire du vocabulaire scientifique 7 (1985) 77–95.

[6] A. Prost, Les mots, in: R. René (Ed.), Pour une histoire politique, Seuil, Paris, 1988, pp. 255–286.

[7] D. Mayaffre, De la lexicométrie à la logométrie, Astrolabe.

[8] Rastier, La mesure et le Grain. Sémantique de corpus, champion Edition, Paris, 2011.

[9] W. Martinez, A. Salem, Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels, Thèse doctorat (2003).

[10] B. Etienne, Qui lemmatise dilemne attise, lexicometrica (2000) 1–19.

[11] D. F. Morrison, Multivariate Statistical Methods, Mac Graw Hill, New-York, 1967.

[12] J.-P. Benzecri, Correspondence Analysis Handbook, Marcel Dekker, New-York, 1992.

[13] L. Lebart, A. Morineau, K. Warwicki, Multivariate Descriptive Statistical Analysis Correspondence Analysis and Related Techniques for Large Matrices, Wiley, New-York, 1984.

[14] E. Oja, S. Kaski, Kohonen Maps, Elsevier, 1999.

[15] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive text document collection, in: E. Oja, S. Kaski (Eds.), Kohonen Maps, Elsevier, 1999, pp. 171–182.

[16] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, Websom for textual data mining, Artificial Intelligence Review 13 (1999) 345–364.

[17] K. Lagus, S. Kaski, T. Kohonen, Mining massive document collections by the websom method, Information Sciences 163 (2004) 135–156.

[18] A. Rauber, D. Merkl, Text mining in the somlib digital library system: the representations of topics and genres, Applied Intelligence 18 (2003) 271–293.

[19] H. Yang, C. Lee, A text mining approach on automatic generation of web directories and hierarchies, Expert Systems with Applications 27 (2004) 645–663.

[20] N. Janasik, T. Honkela, H. Bruun, Text mining in qualitative research: application of an unsupervised learning methode, Organizational Research Methods 3 (12) (2009) 436–460.

[21] M. Cottrell, J.-C. Fort, G. Pagès, Theoretical aspects of the SOM algorithm, Neurocomputing 21 (1998) 119–138.

[22] M. Cottrell, P. Letrémy, How to use the kohonen algorithm to simultaneously analyze individuals and modalities in a survey, Neurocomputing 63 (2005) 193–207.

[23] T. Kohonen, Self-Organizing Maps, Vol. 30, Springer Series in Information Science, Berlin, 1995.

[24] E. de Bodt, M. Cottrell, M. Verleysen, Statistical tools to assess the reliability of self-organizing maps, Neural Networks 15, 8-9 (2002) 967–978.

[25] C. Aggarwal, V. E. Lee, N. Ruan, R. Jin, A survey of algorithms for dense subgraph discovery, Managing and Mining Graph Data, Advances in Database Systems 40 (2010) 303–336.

[26] N. Bourgeois, G. Giannakos, I. Milis, V. T. Paschos, O. Pottie, The max quasi-independent set problem, Journal of Combinatorial Optimization 23 (2010) 94–117.

[27] N. Bourgeois, B. Escoffier, V. T. Paschos, J. M. M. v. Rooij, Fast algorithms for max independent set, Algorithmica 62 (2012) 382–415.

[28] B. Deruelle, Enjeux politiques et sociaux de la culture chevaleresque au XVIe siècle: les prologues de chansons de geste imprimées, Revue historique (3) (2010) 551–576.